Trial Monitoring:
Statistical Challenges and Multiple Outcomes

Garnet Anderson, PhD
WHI Clinical Coordinating Center
Fred Hutchinson Cancer Research Center

June 24, 2003

Overview of Trial Monitoring

- Background
- Procedural aspects
- Statistical challenges
- Developing the WHI monitoring plan

Trial Monitoring Background

Purpose of trial monitoring

- Assure the ethical conduct of the trial
  - Limit exposure to clearly inferior treatments
  - Avoid unnecessary experimentation
  - Assure appropriate steps are taken to ameliorate risk
- Assure that results will be valid and credible

Who monitors clinical trials?

- Investigators
- Sponsor
- Data and Safety Monitoring Boards
  - Membership: scientists, physicians, consumers, ethicists
  - Selected for:
    - Expertise relevant to trial hypotheses
    - Skills in assessing data
    - Perspective on relevant health issues
    - Freedom from “conflict of interest”

Scope of trial monitoring

- Design and consent
- Recruitment
- Adherence
- Outcomes assessment
- Data quality
- Intervention effects on outcomes

Scope of monitoring

- Some trials may need more limited monitoring
  - Low-powered studies
  - Intermediate outcome trials
  - Unbiased interim data cannot be obtained
  - Long interval between intervention and outcome

Pocock SJ. Clinical Trials: A practical approach. Wiley, 1983

Prevention trials features that affect monitoring

- Ostensibly healthy participants
- Low morbidity and mortality rates
- Interventions may have effects on several diseases
- Unlikely to be repeated

Statistical challenges in monitoring prevention trials

- Incorporating multiple endpoints including endpoint-specific
  - Incidence rates
  - Disease burden
  - Size of intervention effects
  - Lag time to intervention effects

Green and Freedman (1994) Statistics in Medicine

Procedural aspects of trial monitoring

Investigator responsibilities

- Propose a trial monitoring plan aligned with
  - Motivating hypotheses
  - Strengths of the trial design and implementation
- Collect, analyze and report data
  - Analysis and reporting should be limited to investigators without participant contact

DSMB responsibilities

- Review accumulating data
- Assure participant safety
- Assess treatment efficacy

Wittes (1993) Statistics in Medicine

Review accumulating data

- Achieving recruitment goals
- Adherence to protocol
  - Eligibility
  - Interventions
  - Data collection
- Data quality

Assure participant safety

- Examine pre-specified safety endpoints
- Consider possible unanticipated intervention effects

Assess treatment efficacy

- Limit monitoring to pre-specified endpoints
- Avoid over-reliance on intermediate endpoints
- Determine if stated hypotheses have been adequately tested
  - Clear evidence of intervention effect
  - Convincing evidence of no effect

Other monitoring considerations

- Data preparation
  - Need an unbiased picture of the data
- Frequency of interim analyses
- Confidentiality
- Blinding of DSMB
- Delineation of responsibilities for decisions
- An early stopping plan

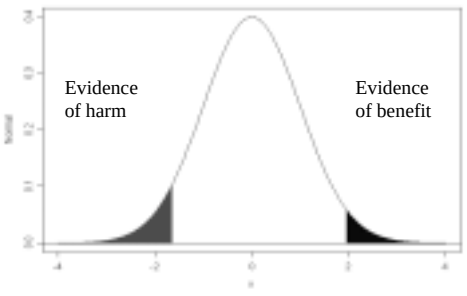
Pocock SJ. Clinical Trials: A practical approach. Wiley, 1983

Statistical challenges in monitoring

- Accommodating asymmetry in risk and benefit decisions:
  - Allocation of type I error to the two tails
  - Spending function differences

Levels of statistical evidence

$Pr(X < -1.645) = 0.05$ and $Pr(X > 1.96) = 0.025$



Statistical challenges in monitoring

- Avoiding inflation of type I errors associated with multiple outcomes
  - Bonferroni correction—
    - Divide level of test (typically α=0.05) by number of outcomes
    - Or, multiply observed p-value by number of outcomes

  ➢ Easy to implement
  ➢ Applicable to every setting
  ➢ Generally quite conservative, especially for correlated outcomes

Statistical challenges in monitoring

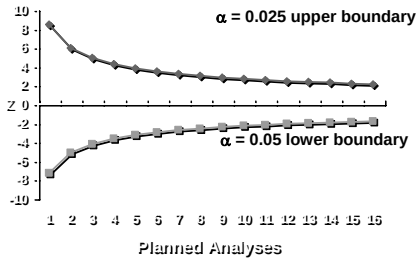
- Avoiding inflation of type I errors associated with multiple 'looks'
  - Group-sequential methods
    - Pocock (1977) Biometrika
    - O'Brien and Fleming (1979) Biometrics
    - Lan and DeMets (1983) Biometrika

Repeated tests on accumulating data

# of repeated 0.05-level tests	Overall significance level
1	0.05
2	0.08
3	0.11
4	0.13
5	0.14
10	0.19
20	0.25
100	0.37

Armitage et al. 1969

O'Brien-Fleming Boundaries



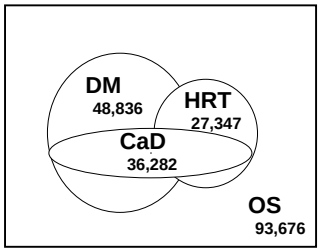
Formal monitoring plan useful for

- Assure statistical properties of procedures
- Avoid over-interpretation of emerging data
- Assist in balancing potential risks and benefits

Developing a monitoring plan

An example from WHI

Design of WHI



CT = 68,133
WHI = 161,809

WHI primary & secondary outcomes

	DM	HRT	CaD
CHD	2°	1°	X
Angina	2°	2°	X
Revascularization	2°	2°	X
CHF	2°	2°	X
Peripheral vascular disease	2°	2°	X
Stroke	2°	2°	X
Venous thromboembolic disease	X	2°	X
Total CVD	2°	2°	X
Breast cancer	1°	2°	2°
Colorectal cancer	1°	X	2°
Endometrial cancer	2°	2°	X
Ovarian Cancer	2°	2°	X
Total Cancer	2°	2°	2°
Hip Fractures	X	2°	1°
Other Fractures	X	2°	2°
Diabetes	2°	X	X
Total Mortality	2°	2°	2°

Specialties represented in WHI DSMB

- Cardiology
- Endocrinology
- Epidemiology
- Gynecology
- Oncology
- Statistics
- Nutrition
- Ethics
- Behavioral Science

Initial DSMB agreement on

- Separate termination decisions for each CT component
- Component-specific list of endpoints
- Use of protocol-defined weighted logrank statistics
- No adjustment for multiple CT components
- Need mechanism to monitor unanticipated effects
- Use of O'Brien-Fleming group sequential methods
- Asymmetry of risks and benefits

Jointly monitoring risks and benefits

- Needed a 'global index' that
  - Provided a quantitative assessment of risks and benefits
  - Would be tailored to hypothesized effects
  - Could play a leading or supportive role

Purely global approaches

- Total mortality
  - Advantage: A compelling endpoint
  - Disadvantage: Limited sensitivity
- Total morbidity
  - Advantage: Sensitive
  - Disadvantages: Problems in definition and ascertainment

Combined index definition

A combined index of endpoint effects can be written as

$$U = S\, w_i\, d_i$$

where:

d_i = observed difference in proportions for the ith endpoint

w_i = weight associated with the ith endpoint

Combined index options

- Possible elements of index
  - Primary only
  - Secondary and safety endpoints
  - Death from other causes
- Choice of weights
  - Expected proportion of deaths
  - Expected years of life lost
  - Quality of life
  - Bayesian priors according to the level of preliminary evidence of effect

Scenario 2-DM

6 years of average follow-up

	C (N=28,800)		I (N=19,200)		
	%	SE	%	SE	Z
Incidence					
Breast Cancer	2.05	0.08	1.85	0.10	1.56
Colorectal Cancer	1.07	0.06	0.92	0.07	1.63
CHD	3.02	0.10	2.63	0.12	2.54*
Mortality					
Breast Cancer	0.51	0.04	.046	0.05	0.78
Colorectal Cancer	0.37	0.04	0.32	0.04	0.97
CHD	1.21	0.06	1.05	0.07	1.64
Other causes	5.50	0.13	5.11	0.16	1.85

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming

Results for Scenario 2-DM

- DSMB opinions
  - 8 continue, 2 stop, 2 cannot decide — Continue
- Statistical methods
  - Primary outcomes — Continue
  - Global methods
    - Total mortality — Stop
    - Unweighted combination — Stop
    - Weighted combination — Stop
    - Bayesian weighted combination — Stop
  - Mixed Methods
    - 1o + global index significant — Continue
    - 1o + global index supportive — Continue

Scenario 3-DM

	6 years of average follow-up				
	C (N=28,800)		I (N=19,200)		
	%	SE	%	SE	Z
Incidence					
Breast Cancer	2.05	0.08	1.72	0.09	2.63*
Colorectal Cancer	1.07	0.06	0.83	0.07	2.69*
CHD	3.02	0.10	3.02	0.12	0.00
Mortality					
Breast Cancer	0.51	0.04	0.43	0.05	1.27
Colorectal Cancer	0.37	0.04	0.29	0.04	1.59
CHD	1.21	0.06	1.21	0.08	0.00
Other causes	5.50	0.13	5.50	0.16	0.00

***Exceeds the 5% critical level of 2.45 using O'Brien and Fleming**

Results for Scenario 3-DM

- DSMB opinions
  - 3 continue, 7 stop, 2 cannot decide — Stop(?)
- Statistical methods
  - Primary outcomes — Stop
  - Global methods
    - Total mortality — Continue
    - Unweighted combination — Continue
    - Weighted combination — Continue
    - Bayesian weighted combination — Continue
  - Mixed Methods
    - 1o + global index significant — Continue
    - 1o + global index supportive — Stop

Scenario 4-HRT/ERT

6 years of average follow-up

	C (N=10,500+)		I (N=7,500+)		
	%	SE	%	SE	Z
Incidence					
CHD	3.26	0.17	2.59	0.18	2.66*
Hip Fractures	1.87	0.13	1.37	0.13	2.65*
Breast Cancer	2.07	0.14	2.25	0.17	-0.82
Endometrial Cancer+	0.46	0.07	1.30	0.13	-5.72*
Mortality					
CHD	1.30	0.11	1.04	0.12	1.61
Hip Fractures	0.47	0.07	0.34	0.07	1.37
Breast Cancer	0.52	0.07	0.56	0.09	-0.36
Endometrial Cancer+	0.05	0.02	0.13	0.04	-1.80
Other causes	5.37	0.22	5.37	0.26	0.00

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming
+Based on initial protocol

Results for Scenario 4-HRT/ERT

- DSMB opinions
  - 6 continue, 5 stop, 1 cannot decide — Continue(?)
- Statistical methods
  - Primary outcomes — Stop
  - Global methods
    - Total mortality — Continue
    - Unweighted combination — Continue
    - Weighted combination — Continue
    - Bayesian weighted combination — Continue
  - Mixed Methods
    - 1o + global index significant — Continue
    - 1o + global index supportive — Continue
    - 1o/adverse effect + global index supportive — Continue

Scenario 6-CaD

6 years of average follow-up

	C (N=22,500)		I (N=22,500)		
	%	SE	%	SE	Z
Incidence					
Hip Fractures	1.51	0.08	1.21	0.07	2.75*
Colorectal Cancer	0.86	0.06	0.75	0.06	1.31
Mortality					
Hip Fractures	0.38	0.04	0.30	0.04	1.46
Colorectal Cancer	0.30	0.04	0.26	0.03	1.02
Other causes	5.92	0.16	5.86	0.16	0.27

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming

Results for Scenario 6-CaD

- DSMB opinions
  - 3 continue, 7 stop, 2 cannot decide — Stop (?)
- Statistical methods
  - Primary outcomes — Stop
  - Global methods
    - Total mortality — Continue
    - Unweighted combination — Continue
    - Weighted combination — Continue
    - Bayesian weighted combination — Continue
  - Mixed Methods
    - 1o + global index significant — Continue
    - 1o + global index supportive — Stop

Scenario 7-HRT/ERT

6 years of average follow-up

	C (N=10,500+)		I (N=7,500+)		
	%	SE	%	SE	Z
Incidence					
CHD	3.26	0.17	3.04	0.20	0.84
Hip Fractures	1.87	0.13	1.74	0.15	0.65
Breast Cancer	2.07	0.14	2.43	0.18	-1.60
Endometrial Cancer+	0.46	0.07	1.30	0.13	5.72
Mortality					
CHD	1.30	0.11	1.22	0.13	0.48
Hip Fractures	0.47	0.07	0.44	0.08	0.30
Breast Cancer	0.52	0.07	0.61	0.09	-0.79
Endometrial Cancer+	0.05	0.02	0.13	0.04	-1.80
Other causes	5.37	0.22	5.37	0.26	0.00

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming
+Based on initial protocol

Results for Scenario 7-HRT/ERT

- DSMB opinions
  - 3 continue, 5.5 stop, 3.5 cannot decide! — Stop (?)
- Statistical methods
  - Primary outcomes — Continue
  - Global methods
    - Total mortality — Continue
    - Unweighted combination — Continue
    - Weighted combination — Continue
    - Bayesian weighted combination — Continue
  - Mixed Methods
    - 1o + global index significant — Continue
    - 1o + global index supportive — Continue
    - 1o/adverse effect + global index supportive — Stop

Scenario 8-HRT/PERT

- 8 years of average follow-up

	C (N=6,500+)		I (N=7,000+)		
	%	SE	%	SE	Z
Incidence					
CHD	3.26	0.23	3.04	0.21	0.72
Hip Fractures	1.87	0.17	1.74	0.16	0.56
Breast Cancer	2.07	0.18	2.79	0.20	-2.69*
Endometrial Cancer	0.46	0.09	0.46	0.08	0.00
Mortality					
CHD	1.30	0.14	1.22	0.13	0.41
Hip Fractures	0.47	0.09	0.44	0.08	0.25
Breast Cancer	0.52	0.09	0.70	0.10	-1.33
Endometrial Cancer	0.05	0.03	0.05	0.03	0.00
Other causes	5.37	0.29	5.37	0.28	0.00

*Exceeds the 5% critical level of 2.45 using O'Brien and Fleming
+Based on initial protocol

Results for Scenario 8-HRT/PERT

- DSMB opinions
  - 0 continue, 12 stop, 0 cannot decide — Stop
- Statistical methods
  - Primary outcomes — Continue
  - Global methods
    - Total mortality — Continue
    - Unweighted combination — Continue
    - Weighted combination — Continue
    - Bayesian weighted combination — Continue
  - Mixed Methods
    - 1o + global index significant — Continue
    - 1o + global index supportive — Continue
    - 1o/adverse effect + global index supportive — Stop

Summary of scenario results

	1	2	3	4	5	6	7	8
DSMB majority opinion	C	C	S?	C?	C	S?	S?	S
Primary endpoint	C	C	S	S	C	S	C	C
Global methods								
Total mortality	C	S	C	C	C	C	C	C
Unweighted combination	C	S	C	C	C	C	C	C
Weighted combination	C	S	C	C	C	C	C	C
Bayesian weighted	C	S	C	C	C	C	C	C
Mixed methods								
1^o + global significant	C	C	C	C	C	C	C	C
1^o + global supportive	C	C	S	C	C	S	C	C
1^o or adverse effect + global significant	C	C	S	S	C	S	S	S

Conclusions from exercise

- Monitoring primary endpoint was insufficient
- Global indices
  - Performed similarly
  - Were somewhat insensitive to overall balance of risks and benefits
- Mixed approach using primary endpoint supported by a global index best captured DSMB consensus

Conclusions from exercise

- Needed more sensitivity to pre-specified adverse effects
- Use of scenarios was very beneficial to creating formal monitoring plan

WHI monitoring plan for E+P trial

A Case Study in Early Stopping

E+P monitoring plan

- Primary Endpoint: CHD
- Primary Safety Endpoint: Breast Cancer
- Secondary Endpoints:
  - Hip fractures
  - Stroke
  - Pulmonary Embolism
  - Endometrial Cancer
  - Colorectal Cancer
  - Death from other causes

WHI Estrogen+Progestin Trial Global Index

- Defined for each woman as the earliest of:
  - CHD
  - Invasive breast cancer
  - Stroke
  - PE
  - Endometrial cancer
  - Colorectal cancer
  - Hip fracture
  - Death from other causes

E+P trial monitoring for benefit Early stopping considerations required:

- Evidence of CHD benefit
  - Statistical rules based on O'Brien-Fleming (OBF) procedures using a 0.025-level, one-sided test
- AND
- Global index supportive of benefit
  - Statistical rules based on OBF procedures using a 0.05-level, one-sided test

O'Brien PC, Fleming TR. Biometrics. 1979;35:549-556.

Trial monitoring for adverse effects Early stopping considerations required:

- Evidence of increase in breast cancer
  - OBF procedure using a 0.05-level one-sided, weighted logrank test.

OR

- Evidence of increase in any of the other 7 pre-specified endpoints
  - OBF procedure using a 0.05-level one-sided, weighted logrank test, with Bonferroni correction.

AND

- Global index supportive of overall harm ($Z < -1.0$)

Freedman, et al. Control Clin Trials. 1996;17:509-525.

Limitations of a monitoring plan

- Real data are more complex than the scenarios
- Care is needed in considering any modification to monitoring plan based on emerging trial data
  - Avoid redefinition of endpoints
- Assumptions underlying the trial design and monitoring plan may be incorrect

Monitoring plan is a guideline

- Emerging external data may impact assessment
- Statistical boundaries provide tools for assessing strength of the data
- Good judgment is always required

Next: Stopping the WHI E+P Trial

The finale of our case study in trial monitoring